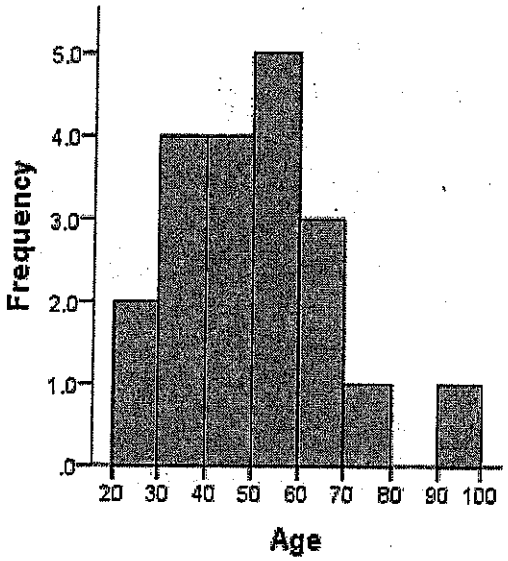
Name: _____ Date: _____

Histograms

Read ↓

5
What is a histogram?

A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g. normal distribution), outliers, skewness, etc. An example of a histogram, and the raw data it was constructed from, is shown below:



36	25	38	46	55	68	72	55	36	38
67	45	22	48	91	46	52	61	58	55

How do you construct a histogram from a continuous variable?

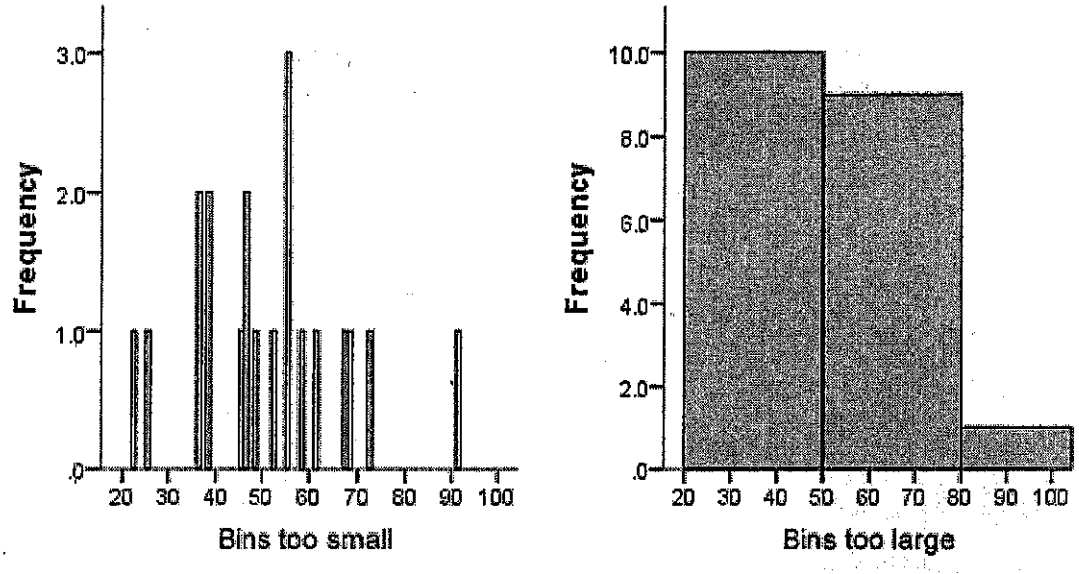
To construct a histogram from a continuous variable you first need to split the data into intervals, called **bins**. In the example above, **age** has been split into bins, with each bin representing a 10-year period starting at 20 years. Each bin contains the number of occurrences of scores in the data set that are contained within that bin. For the above data set, the frequencies in each bin have been tabulated along with the scores that contributed to the frequency in each bin (see below):

Bin	Frequency	Scores Included in Bin
20-30	2	25,22
30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91

Notice that, unlike a bar chart, there are no "gaps" between the bars (although some bars might be "absent" reflecting no frequencies). This is because a histogram represents a continuous data set, and as such, there are no gaps in the data. (Although you will have to decide whether you round up or round down scores on the boundaries of bins)

Choosing the correct bin width

There is no right or wrong answer as to how wide a bin should be, but there are rules of thumb. You need to make sure that the bins are not too small or too large. Consider the histogram we produced earlier (see above): the following histograms use the same data but have either much smaller or larger bins, as shown below:



We can see from the histogram on the left, that the bin width is too small as it shows too much individual data and does not allow the underlying pattern (frequency distribution) of the data to be easily seen. At the other end of the scale, is the diagram on the right, where the bins are too large and, again, we are unable to find the underlying trend in the data.

Histograms are based on area not height of bars

In a histogram, it is the area of the bar that indicates the frequency of occurrences for each bin. This means that the height of the bar does not necessarily indicate how many occurrences of scores there were within each individual bin. It is the product of height multiplied by the width of the bin that indicates the frequency of occurrences within that bin. One of the reasons that the height of the bars is often incorrectly assessed as indicating frequency and not the area of the bar is due to the fact that a lot of histograms often have equally spaced bars (bins) and, under these circumstances, the height of the bin does reflect the frequency.

What is the difference between a bar chart and a histogram?

The major difference is that a histogram is only used to plot the frequency of score occurrences in a continuous data set that has been divided into classes, called bins. Bar charts, on the other hand, can be used for a great deal of other types of variables including ordinal and nominal data sets.

DO

Histogram Worksheet Name_____

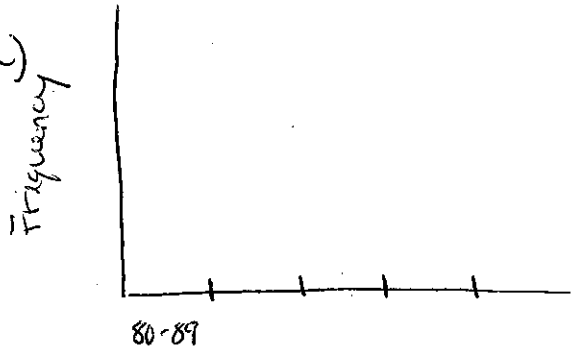
Create a frequency table and Histogram using the given information.

Number of crimes committed in 1984.

January	124	February	96	March	89
April	113	May	107	June	102
July	85	August	87	September	91
October	119	November	122	December	115

Interval	frequency
80 – 89	3
90 - 99	
100 – 109	
110 – 119	
120 - 129	

(July, August, March)

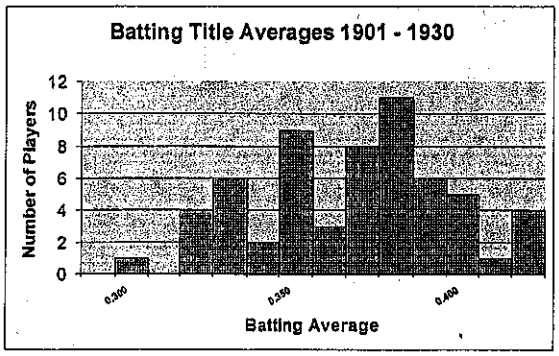


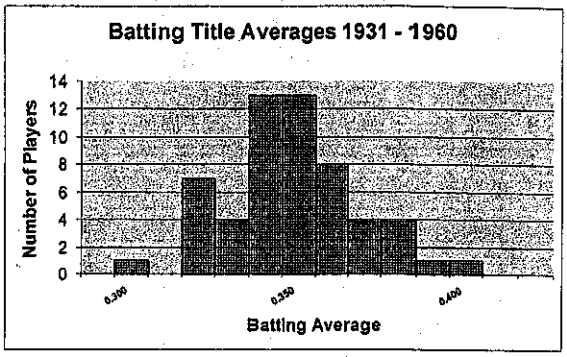
Test scores for a high school Biology Test

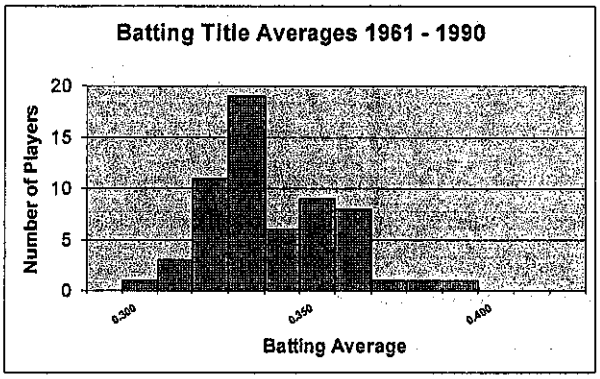
81	77	63	92	97	68	72
88	78	96	85	70	66	95
80	99	63	58	83	93	75
89	94	92	85	76	90	87

Interval	frequency
60 - 69	
70 – 79	
80 - 89	
90 - 99	

The 3 histograms below show the batting averages of the winners of the batting title in the major league baseball (for both the American & National leagues) for certain years in the 1900s. Batting average shows the percent (written as a decimal) of the time a certain player gets a hit. A player who has a batting average of 0.405 has gotten a hit in 40.5 % of the times that they were at bat. The batting title is an award given to the player with the highest batting average for a given season. Refer to the histograms as you answer questions 1 – 6.







_____1. How many batting titles were won with a batting average of between 0.300 and 0.350 from 1901 to 1930?

_____2. How many batting titles were won with a batting average of between 0.300 and 0.350 from 1931 to 1960?

_____3. How many batting titles were won with a batting average of between 0.300 and 0.350 from 1961 to 1990?

4. If you were to find the mean of each of the winning batting averages for each time period, which time period do you think would have the highest mean? Explain.

_____

_____

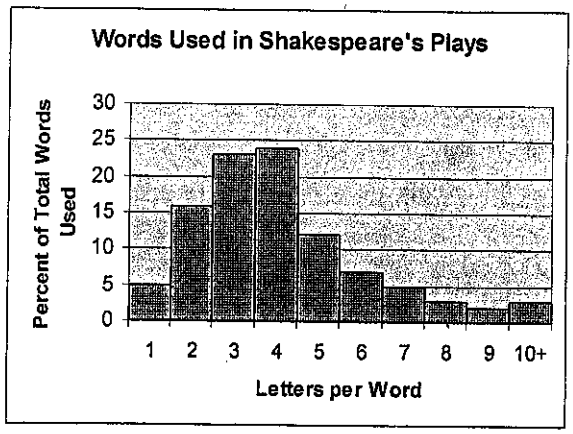
5. As the century progressed, what in general happened to the batting averages of the batting title winners? Explain.

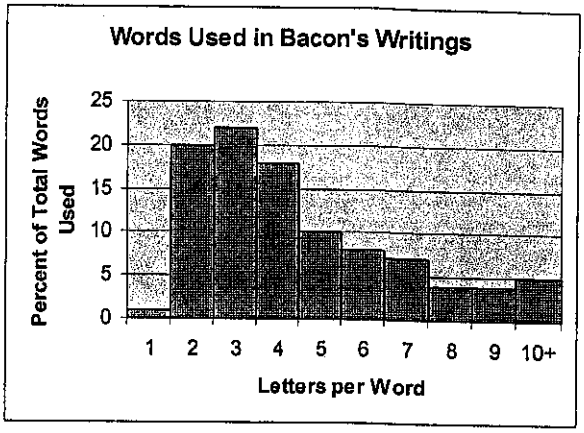
_____

_____

_____

For questions 6 – 10, refer to the following 2 histograms. These histograms were made in an attempt to determine if William Shakespeare was really just a pen name for Sir Francis Bacon. (A pen name is a fake name used by another person when writing). A few scholars have had this idea and in order to determine if this was true, a researcher had to count the letters in every word of Shakespeare's plays & Bacon's writing (and you thought you had a lot of homework). Their results are recorded in the histograms below.





_____6. What percent of all Shakespeare's words are 4 letters long?

_____7. What percent of all Bacon's words are 4 letters long?

_____8. What percent of all Shakespeare's words are more than 5 letters long?

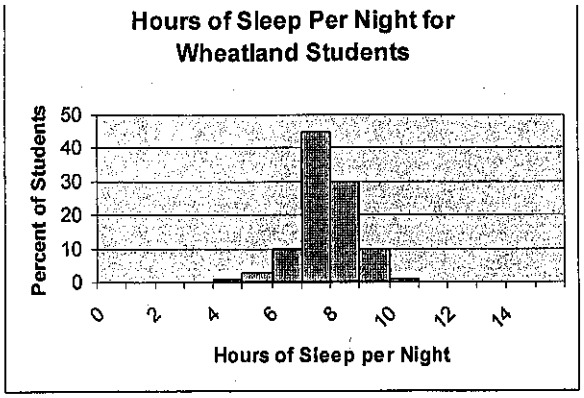
_____9. What percent of all Bacon's words are more than 5 letters long?

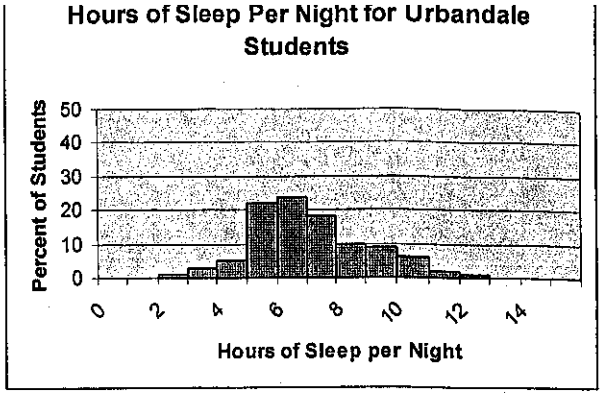
10. Based on these histograms, do you think that William Shakespeare was really just a pen name for Sir Francis Bacon? Explain.

_____

_____

_____





Suppose that the two histograms above show the sleeping habits of the teens at two different high schools. Wheatland High School is a small rural school consisting of 100 students while Urbandale High School is located in a large city and has 3,500 students.

_____11. About what percent of the students at Wheatland get at least 8 hours of sleep per night?

_____12. About what percent of the students at Urbandale get at least 8 hours of sleep per night?

_____13. Which high school has more actual students that sleep between 9 – 10 hours per night?

_____14. Which high school has a higher median sleep time?

15. Wheatland's percent of students who sleep between 8-9 hours a night is _____ % more than Urbandale's percent of students who sleep between 8-9 hours per night.

16. Consider the type of data in the last two sets of problems (letters per word & sleep times).

_____a) Are letters per word qualitative or quantitative?

_____b) Are sleep times qualitative or quantitative?

_____c) Which data set is continuous?

_____d) Which data set is discrete